Notes on Matrix Derivative in Neural Network

Wu Xiaojian

August 3rd, 2019

1 Notation

Let lower case letters $a, b, c, \cdots$ be scalar values, bold lower case letters $\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \cdots$ be vectors and bold upper case letters $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \cdots$ be matrices. Detailedly, we have

$$
\begin{aligned}
a, b, c, \cdots &\in \mathbb{R} \\
\boldsymbol{v} = [v_1, v_2, \cdots, v_n]^T &\quad \text{for } \boldsymbol{v} \in \mathbb{R}^n \\
\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} &\quad \text{for } \boldsymbol{A} \in \mathbb{R}^{m \times n}
\end{aligned}
$$

We introduce a notation $\delta_{ij,pq}$, which will be used in the formulas later, as below:

$$
\delta_{ij,pq} = \begin{cases} 1 & \text{if } i = p \text{ and } j = q \\ 0 & \text{otherwise} \end{cases}
$$

in some cases, we are to leave out one term and write $\delta_{i,p}$, meaning 1 if $i = p$, and 0 otherwise.

2 Derivative of Scalar, Vector and Matrix

Let $L \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we define derivative operations among them:

- scalar value and vector

$$
\frac{\partial L}{\partial \boldsymbol{x}} = \left[\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \cdots, \frac{\partial L}{\partial x_n} \right] \tag{1a}
$$

$$
\frac{\partial \boldsymbol{x}}{\partial L} = \left[\frac{\partial x_1}{\partial L}, \frac{\partial x_2}{\partial L}, \cdots, \frac{\partial x_n}{\partial L} \right]^T \tag{1b}
$$

  NOTICE: The Derivative of L to $\boldsymbol{x}$ is a **row vector** of $1 \times n$ but not a column vector of the same shape of $\boldsymbol{x}$, while the derivative of $\boldsymbol{x}$ to L is a column vector with the shape consistent to $\boldsymbol{x}$.

- vector to vector

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \frac{\partial y_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{2}$$

  we give a name to such a matrice as **Jacobian Matrix**.

- scalar value to matrice

$$\frac{\partial L}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial L}{\partial a_{11}} & \frac{\partial L}{\partial a_{21}} & \cdots & \frac{\partial L}{\partial a_{m1}} \\ \frac{\partial L}{\partial a_{12}} & \frac{\partial L}{\partial a_{22}} & \cdots & \frac{\partial L}{\partial a_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial a_{1n}} & \frac{\partial L}{\partial a_{2n}} & \cdots & \frac{\partial L}{\partial a_{mn}} \end{bmatrix} \tag{3}$$

  Like EQUATION(1a), the result matrice is obtained by applying derivative of L to each entry of $\boldsymbol{A}$ and taking **transpose operation** on it, so that the shape is $n \times m$ but not $m \times n$.

In multi-variable calculas, we know that if $L = L(y_1, y_2, \cdots, y_n)$ and $y_i = y_i(x_1, x_2, \cdots, x_m)$ $(i = 1, 2, \cdots, n)$, then

$$\frac{\partial L}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial x_k}$$

(**NOTICE:** In single-variable derivative, say, $z = f(y), y = g(x)$, by **chain rule** we have $\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}$. But in partial derivative, we cannot conclude that $\frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_k} = \frac{\partial L}{\partial x_k}$)

Let's regard these functions in another way. Since L takes n arguments as input, it's natural to rewrite L as $L = L(\boldsymbol{y}) = L([y_1, y_2, \cdots, y_n])$, where $y_i = y_i(\boldsymbol{x}) = y_i([x_1, x_2, \cdots, x_m])$. Now we call L and y_i **vector functions**. Hence, by EQUATION(1a) and (1b), we have

$$\begin{matrix} \textit{Chain rule for} \\ \textit{vector function} \end{matrix} : \quad \frac{\partial L}{\partial x_k} = \frac{\partial L}{\partial \boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial x_k} \tag{4}$$

3 Derivatives for Matrix in Neural Network

In neural network, we assume the prior layer, say, $(w-1)$-th layer, outputs m data $\boldsymbol{X}$, namely $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(m)}]$ where $\boldsymbol{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}\right]^T$ $(i = 1, 2, \cdots, m$, and $x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}$ are m attributes or features of $\boldsymbol{x}^{(i)})$, to the w-th layer. The w-th layer takes linear operation on $\boldsymbol{X}$ and outputs $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{X} + \boldsymbol{B} \tag{5}$$

where $\boldsymbol{W}$ is weight matrix and $\boldsymbol{B}$ is bias matrix:

$$\boldsymbol{Z} = \left[\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \cdots, \boldsymbol{z}^{(m)}\right] \tag{6a}$$

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pn} \end{bmatrix}, \quad \boldsymbol{B} = \underbrace{\begin{bmatrix} b_1 & b_1 & \cdots & b_1 \\ b_2 & b_2 & \cdots & b_2 \\ \vdots & \vdots & \ddots & \vdots \\ b_p & b_p & \cdots & b_p \end{bmatrix}}_{m \text{ columns}} \tag{6b}$$

We can multiply $\boldsymbol{W}$ to each entry of $\boldsymbol{X}$ and therefore,

$$\boldsymbol{z}^{(k)} = \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{b} \quad (k = 1, 2, \cdots, m) \tag{7}$$

where $\boldsymbol{b} = [b_1, b_2, \cdots, b_p]^T$.

3.1 For single datum

For simplicity, let's first consider the case of only one datum, the k-th example: $\boldsymbol{z}^{(k)} = \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{b}$, namely

$$\begin{bmatrix} z_1^{(k)} \\ z_2^{(k)} \\ \vdots \\ z_p^{(k)} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pn} \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \tag{8}$$

It's obvious that $(i = 1, 2, \cdots, p)$

$$z_i^{(k)} = \sum_{j=1}^{n} w_{ij} x_j^{(k)} + b_i \tag{9}$$

Let $L^{(k)} : \mathbb{R}^p \mapsto \mathbb{R}$ be $L^{(k)} = L^{(k)}(\boldsymbol{z}^{(k)})$. We will evaluate $\frac{\partial L^{(k)}}{\partial \boldsymbol{W}}$, $\frac{\partial L^{(k)}}{\partial \boldsymbol{x}^{(k)}}$ and $\frac{\partial L^{(k)}}{\partial \boldsymbol{b}}$.

By Equation(3), we have

$$\frac{\partial L^{(k)}}{\partial \boldsymbol{W}} = \begin{bmatrix} \frac{\partial L^{(k)}}{\partial w_{11}} & \frac{\partial L^{(k)}}{\partial w_{21}} & \cdots & \frac{\partial L^{(k)}}{\partial w_{p1}} \\ \frac{\partial L^{(k)}}{\partial w_{12}} & \frac{\partial L^{(k)}}{\partial w_{22}} & \cdots & \frac{\partial L^{(k)}}{\partial w_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L^{(k)}}{\partial w_{1n}} & \frac{\partial L^{(k)}}{\partial w_{2n}} & \cdots & \frac{\partial L^{(k)}}{\partial w_{pn}} \end{bmatrix} \tag{10}$$

Now let's take one entry from matrice above, say, $\frac{\partial L^{(k)}}{\partial w_{ij}}$. Since $L^{(k)} = L^{(k)}(\boldsymbol{z}^{(k)})$ and $\boldsymbol{z}^{(k)} = \boldsymbol{z}^{(k)}(\boldsymbol{W})$, by Equation(4), we have that

$$\frac{\partial L^{(k)}}{\partial w_{ij}} = \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \frac{\partial \boldsymbol{z}^{(k)}}{\partial w_{ij}} \tag{11}$$

where

$$\frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} = \left[\frac{\partial L^{(k)}}{\partial z_1^{(k)}}, \frac{\partial L^{(k)}}{\partial z_2^{(k)}}, \cdots, \frac{\partial L^{(k)}}{\partial z_p^{(k)}}\right] \tag{12a}$$

$$\frac{\partial \boldsymbol{z}^{(k)}}{\partial w_{ij}} = \left[\frac{\partial z_1^{(k)}}{\partial w_{ij}}, \frac{\partial z_2^{(k)}}{\partial w_{ij}}, \cdots, \frac{\partial z_p^{(k)}}{\partial w_{ij}}\right]^T \tag{12b}$$

Using EQUATION(9), there is, for $h = 1, 2, \cdots, p$

$$\begin{aligned}
\frac{\partial z_h^{(k)}}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}}\left(\sum_{t=1}^{n} w_{ht} x_t^{(k)} + b_h\right) \\
&= \sum_{t=1}^{n} \frac{\partial w_{ht}}{\partial w_{ij}} x_t^{(k)} \\
&= \sum_{t=1}^{n} \delta_{ij,ht} x_t^{(k)} \\
&= \delta_{i,h} x_j^{(k)}
\end{aligned} \tag{13}$$

This infers that the i-th entry of $\frac{\partial \boldsymbol{z}^{(k)}}{\partial w_{ij}}$ is $x_j^{(k)}$ while the others are 0, and so we conclude that

$$\frac{\partial L^{(k)}}{\partial w_{ij}} = \frac{\partial L^{(k)}}{\partial z_i^{(k)}} x_j^{(k)} \tag{14}$$

Hence,

$$\begin{aligned}
\frac{\partial L^{(k)}}{\partial \boldsymbol{W}} &= \begin{bmatrix}
\frac{\partial L^{(k)}}{\partial z_1^{(k)}} x_1^{(k)} & \frac{\partial L^{(k)}}{\partial z_2^{(k)}} x_1^{(k)} & \cdots & \frac{\partial L^{(k)}}{\partial z_p^{(k)}} x_1^{(k)} \\
\frac{\partial L^{(k)}}{\partial z_1^{(k)}} x_2^{(k)} & \frac{\partial L^{(k)}}{\partial z_2^{(k)}} x_2^{(k)} & \cdots & \frac{\partial L^{(k)}}{\partial z_p^{(k)}} x_2^{(k)} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial L^{(k)}}{\partial z_1^{(k)}} x_n^{(k)} & \frac{\partial L^{(k)}}{\partial z_2^{(k)}} x_n^{(k)} & \cdots & \frac{\partial L^{(k)}}{\partial z_p^{(k)}} x_n^{(k)}
\end{bmatrix} \\
&= \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} \begin{bmatrix} \frac{\partial L^{(k)}}{\partial z_1^{(k)}} & \frac{\partial L^{(k)}}{\partial z_2^{(k)}} & \cdots & \frac{\partial L^{(k)}}{\partial z_p^{(k)}} \end{bmatrix} \\
&= \boldsymbol{x}^{(k)} \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}}
\end{aligned} \tag{15}$$

Since the derivative of scalar value w.r.t. a matrice is still a matrice whose size is as the same as that of the transpose of the origin one, it's sometimes very convenient in mechine learning to have the derivative matrice size invariant. Here, we define

$$\nabla_{\boldsymbol{W}} L^{(k)} = \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{W}}\right)^T = \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}}\right)^T \left(\boldsymbol{x}^{(k)}\right)^T \tag{16}$$

Now let's calculate $\frac{\partial L^{(k)}}{\partial \boldsymbol{b}}$. Similarly, for $i = 1, 2, \cdots, p$, we have

$$\begin{aligned}
\frac{\partial L^{(k)}}{\partial b_i} &= \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \frac{\partial \boldsymbol{z}^{(k)}}{\partial b_i} \\
&= \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \left[\frac{\partial z_h^{(k)}}{\partial b_i} \right]_{(h=1,2,\cdots,p)}^{p\times 1} \\
&= \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \left[\frac{\partial}{\partial b_i} \left(\sum_{t=1}^{n} w_{ht} x_t^{(k)} + b_h \right) \right]_{(h=1,2,\cdots,p)}^{p\times 1} \\
&= \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \left[\delta_{h,i} \right]_{(h=1,2,\cdots,p)}^{p\times 1} \\
&= \frac{\partial L^{(k)}}{\partial z_i^{(k)}} \quad \left(\begin{array}{l} \textit{The } i\textit{-th component of vector is } 1\textit{, while} \\ \textit{the others are } 0\textit{'s.} \end{array} \right)
\end{aligned} \tag{17}$$

and thus

$$\frac{\partial L^{(k)}}{\partial \boldsymbol{b}} = \left[\frac{\partial L^{(k)}}{\partial z_1^{(k)}}, \frac{\partial L^{(k)}}{\partial z_2^{(k)}}, \cdots, \frac{\partial L^{(k)}}{\partial z_p^{(k)}} \right] = \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \tag{18}$$

we define

$$\nabla_{\boldsymbol{b}} L^{(k)} = \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{b}} \right)^T = \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \right)^T \tag{19}$$

Finally, let's look at $\frac{\partial L^{(k)}}{\partial \boldsymbol{x}^{(k)}}$. Firstly, according to chain rule, we have

$$\frac{\partial L^{(k)}}{\partial \boldsymbol{x}^{(k)}} = \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \frac{\partial \boldsymbol{z}^{(k)}}{\partial \boldsymbol{x}^{(k)}} \tag{20}$$

by Equation(2),

$$\begin{aligned}
\frac{\partial \boldsymbol{z}^{(k)}}{\partial \boldsymbol{x}^{(k)}} &= \left[\frac{\partial z_i^{(k)}}{\partial x_j^{(k)}} \right]_{ij}^{p\times n} \\
&= \left[\frac{\partial}{\partial x_j^{(k)}} \left(\sum_{t=1}^{n} w_{it} x_t^{(k)} + b_i \right) \right]_{ij}^{p\times n} \\
&= \left[\sum_{t=1}^{n} w_{it} \delta_{t,j} \right]_{ij}^{p\times n} \\
&= \left[w_{ij} \right]_{ij}^{p\times n} = \boldsymbol{W}
\end{aligned} \tag{21}$$

What a elegant result it is! We can rewrite this formula as

$$\frac{\partial \boldsymbol{z}^{(k)}}{\partial \boldsymbol{x}^{(k)}} = \frac{\partial \left(\boldsymbol{W} \boldsymbol{x}^{(k)} + \boldsymbol{b} \right)}{\partial \boldsymbol{x}^{(k)}} = \boldsymbol{W} \tag{22}$$

For digression, this is far analgous to single-variable derivative of linear function $y = kx + b$, which is ${}^{\mathrm{d}y}\!/_{\mathrm{d}x} = k$.

Now back to our topic, we gat

$$\frac{\partial L^{(k)}}{\partial \boldsymbol{x}^{(k)}} = \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \boldsymbol{W} \tag{23}$$

and define

$$\nabla_{\boldsymbol{x}^{(k)}} L^{(k)} = \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{x}^{(k)}} \right)^T = \boldsymbol{W}^T \left(\frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \right)^T \tag{24}$$

3.2 For m data

At present, let's consider a batch of samples together assembled in matrix form

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{X} + \boldsymbol{B} \tag{25}$$

Suppose there are m samples, the m columns of $\boldsymbol{Z}$, and let $L = f(\boldsymbol{L}) = f(L^{(1)}, L^{(2)}, \cdots, L^{(m)})$ where $L^{(k)} = L^{(k)}(\boldsymbol{z}^{(k)})$. Up to now, we just put a simple L definition into consideration, which we define as

$$L = \frac{1}{m} \sum_{k=1}^{m} L^{(k)} \tag{26}$$

Evalute $\frac{\partial L}{\partial \boldsymbol{W}}$, $\frac{\partial L}{\partial \boldsymbol{X}}$ respectively.

By EQUATION(15), we have

$$\frac{\partial L}{\partial \boldsymbol{W}} = \frac{1}{m} \sum_{k=1}^{m} \frac{\partial L^{(k)}}{\partial \boldsymbol{W}} = \frac{1}{m} \sum_{k=1}^{m} \boldsymbol{x}^{(k)} \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} \tag{27}$$

Beacuse the (i,j)-entry of $\boldsymbol{x}^{(k)} \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}}$ is $\frac{\partial L^{(k)}}{\partial z_j^{(k)}} x_i^{(k)}$, the sum of these m matrices turns out to be

$$\begin{aligned} \sum_{k=1}^{m} \boldsymbol{x}^{(k)} \frac{\partial L^{(k)}}{\partial \boldsymbol{z}^{(k)}} &= \left[\sum_{k=1}^{m} \frac{\partial L^{(k)}}{\partial z_j^{(k)}} x_i^{(k)} \right]_{ij}^{n \times p} \\ &= \left[x_i^{(k)} \right]_{ik}^{n \times m} \left[\frac{\partial L^{(k)}}{\partial z_j^{(k)}} \right]_{kj}^{m \times p} \\ &= \boldsymbol{X} \frac{\partial L}{\partial \boldsymbol{Z}} \end{aligned} \tag{28}$$

so we have

$$\nabla_{\boldsymbol{W}} L = \left(\frac{\partial L}{\partial \boldsymbol{W}} \right)^T = \frac{1}{m} \left(\frac{\partial L}{\partial \boldsymbol{Z}} \right)^T \boldsymbol{X}^T = \frac{1}{m} \left(\nabla_{\boldsymbol{Z}} L \right) \boldsymbol{X}^T \tag{29}$$

Similarly, we have

$$\nabla_{\boldsymbol{X}} L = \frac{1}{m} \boldsymbol{W}^T \left(\nabla_{\boldsymbol{Z}} L \right) \tag{30}$$